SAND98-0741A

Distributed Search by an Agent Collective[1]

Laurence R. Phillips, Shannon V. Spires, and Steven Y. Goldsmith .
{lrphill, svspire, sygolds}@sandia.gov
Advance Information Systems Laboratory
Sandia National Laboratories Albuquerque, New Mexico 87185

Abstract We are investigating semantically motivated, agent-facilitated methods for information search and retrieval on the world-wide web. We expect each agent to have a private corpus o information that it shares with other similarly-motivated agents in exchange for information it needs. Structure of the agent corpus is motivated by conceptual structures and ontology formation. We have developed several base technologies for meta-reasoning, moving objects on a net, parsing web structures, and standard agent behavior. Sharing strategies and ontology formation are the current research to its.

Introduction and situation

We are investigating semantically motivated, agent-facilitated methods for information search and retrieval on the world-wide web. We tend to divide the search universe into two halves: The weak / shallow / syntactic / statistical side and the strong / deep / semantic / cognitive side. We perceive the state of large-scale free-text search to be:

1.Weak methods have not adequately exploited but are ultimately limited. [Balabanovic et al.] and [trong et al.] are supported by classical learning and suggest that improveme accuracy and coverage will not be so difficult to achieve (although perhaps not at the same time, according to [Armstrong et al.]). Ultimately, however, it's hard to differentiate among {"This page is about elephants", "This page is about 'Elephant' ", "This page is not about elephants", "page abotu si elepahnts is"} if you can use only weak methods. And, as [Etzioni] points out, "a is the labeling problem: data is abundant ... but it is unlabeled."

2.The gulf between weak ods is a continuum. [Borgo et al.] and [Salton et al.], among o as weak methods begin to exploit large lexicons, high connective enables high accuracy with few type II errors; in essence we are finding e is about." Certainly this is what Cycorp is banking on. on into a structured interlingua [Van Baalen and Fikes] wou what we found into "our" terms and form conceptual strut iece of text "says."

3 Agents are useful in searc hing the web or other large, heterogeneous information structures. They can work at the server to reduce network traffic, subdivide search tasks for load balancing, form and maintain internal knowledge structures to represent what they've found, exchange information with one another and with humans , and

[1] Sandia is a multiprogram laboratory o erated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

work unattended for long periods ([Woolridge and Jennings], [Malone et al.], [Lieberman and Maulsby], [Maes], and others).

4. The core issue is representation and exchange of information among distributed search agents.

Current work

We are developing search agents that can exploit the syntactic, structural content of web pages more fully than existing search engines, exchange information with other agents, and build lexicons of what they find using conceptual structures [Sows], summarization [Rau et al.], [Barzilay and Elhadad], and ontology formation [Guarino], [Gruber], [Uschold and King].

Agents bring some tactical advantages to the system; they can reside locally at the servers to reduce network traffic, they can collaborate in subdividing the search space for load leveling, and they can perform their search tasks in an ongoing manner, continually accreting a corpus of information relevant to a given subject. Such corpi can continually grow as new information becomes available, shed no-longer- relevant information, and can dynamically repackage information gathered from multiple static pages, creating a correct "projection" of the information as required by the current requestor.

Our agents at present can delegate work to other agents, parse a web page into a complex object with compositional semantics, extract and follow links, accept and reject goals, transport across the network from server to server, and communicate with other agents and humans via web pages and e-mail (we are currently integrating KQML into the agent's communication repertoire). We are investigating information exchange mechanisms and delegation strategies.

These search systems operate by starting at one or more root nodes (perhaps as returned by a conventional keyword-based search engine) and crawl the multiple trees to a depth determined a *priori* or dynamically. They can subdivide the search task among several agents. Currently findings are scored using weak metrics. We are investigating formation of conceptual structures so that each agent can reorganize and integrate its findings.

Recent work

Over the past few years, we have developed several support technologies as research vehicles, proof of concept, and, finally, platforms for further work:

1. Distributed CLOS (DCLOS) - An open implementation distributed object system that supports proxies, copies, replicants, remote method invocation, object identity, and other capabilities.

2. SpireStore Object Data Base - A custom object oriented data base, fully integrated with DCLOS, that provides persistence, identity, storage of composites, blobs and

arbitrary data types, versioning, transactions (in the ACID sense), meta-data storage, and many other capabilities.

3 Object Lifecycle Protocol (OLP) - A protocol for management of the entire object lifecycle, fully integrated with DCLOS and SpireStore, that includes an object factory (make-object and validate-object), an object destruction protocol (kill-object), examination and update protocols (view-object and update-object), object collections and iterators, composite objects, protocols for handling unknown and incompletely specified composites, and other capabilities.

4.HTML-to-CLOS-to-HTML Interface (HCHI) - A toolkit for rapid development of complex HTML forms interfaces. Includes a compiler that renders the contents of an HTML stream as a composite CLOS object. Rendered objects are “live” and able to execute class methods, including change-class. This enables the attachment of generators and iterators of object base collections to the compiled tree and subsequent re-rendering as an instantiated web page, providing a powerful development capability for object-oriented, WWW-based data systems.

5.Dynamic Object Server Architecture (DOSA) - A development environment based on all of the above-mentioned technologies that provides frameworks for object-based client / server applications. This technology is a complete solution to the corporate intranet design problem.

6.The Standard Agent Framework – A class hierarchy that provides normative agent behavior for accepting and rejecting goals, op-crating on goal structures, communicating with other agents, delegation, negotiation, and elicitation. The framework was used to implement the Border Trade Facilitation System (Goldsmith, S. Y.; Phillips, L. R., and Spires, S. V. *A Multi-agent System for Coordinating International Shipping*, to be published at Agents ’98, Minneapolis, Minnesota, July 1998)

Conclusion

We have much work to do but we are moving towards a powerful distributed search technology that contains highly connected and dynamic knowledge structure about what’s on the web and can communicated the contents of that structure to human collaborators.

References

[Armstrong et al.] Armstrong, R.; Freitag, D.; Joachims, T., and Mitchell, T. WebWatcher: A Learning Apprentice for the World Wide Web, in the *AAAI Spring Symposium on Info. Gathering from Heterogeneous, Distributed Environments*, March 1995.

[Balabanovic et al.] Balabanovic, M.; Shoham, Y.; and Yun, Y. An adaptive agent for automated web browsing, Stanford Digital Library Project SIDL-WP-1995-0023

[Barzilay and Elhadad] Barzilay, R.; and Elhadad, M. Using lexical chains for text summarization (full citation unavailable) [Borgo et al] Borgo, S.; Guarino, N.; Masolo, C., and Vetere, G. Using a large ontology for internet-based retrieval of object-oriented components

[Etzoni] Etzoni, O. The World Wide Web: quagmire or gold mine? in Comm. of the ACM, November 1996

[Guarino] Guarino, N. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction, and Integration, *Summer school on Information Extraction,* to appear ina volume edited by M. T. Pazienza, Springer-Verlag

[Gruber] Gruber, T. (1993). *Toward Principles for the Design of Ontologies Used for Knowledge Sharing*, Knowledge Systems Laboratory KSL 93-04, Stanford University.

[Lieberman and Maulsby] Lieberman, H., and Maulsby, D., "Instructable agents: Software that just keeps getting better," *IBM Systems Journal*, Vol. 35, Nos. 3&4, 1996.

[Maes] Maes, P., "Agents that Reduce Work and Information Overload," *Communications of the ACM 37*(7, July), 31-40, 1994.

[Malone et al.] Malone, T., Grant, K., Turbak, F., Brobst, S., Cohen, M., Intelligent Information-Sharing Systems, *Communications of the ACM 30*(5, May), 390- 402.1994

[Rau et al.] Rau, L.; Jacobs, P.; and Zernik, U. Information extraction and text summarization using linguistic knowledg acquisition, I*nformation Processing and Management*, October, 1989

[Salton et al.] Salton, G.; Singhal., A.; Buckley, C.; nad Mitra, M. Automatic text decomposition using text segments and text themes, Seventh ACM Conf on Hypertext, March 1996

[Sows] Sowa, J. F., *Conceptual Structures: Information Processing in Mind andMachine*, Addison-Wesley, 479p., 1984.

[Uschold and King] Uschold, M.; and King, M. Towards a methodology for building ontologies, Proc. of *Workshop on Basic Ontological Issues in Knowledge Sharing*, AIAI-TR-183, IJCAI-95, July 1995

[Van Baalen and Fikes] Van Baalen, J.; and Fikes, R. *The role of reversible grammars in translating between representation languages*, Knowledge Systems Laboratory, KSL-93-67, November 1993

[Woolridge and Jennings] Wooldridge, M., and Jennings, N., "Intelligent Agents: Theory and Practice, " in *Knowledge Engineering Review 10(*2), 1995.